INSEAD The Business School for the World®

Working Paper
2020/27/TOM
(Revised version of 2020/05/TOM)

Does Immediate Feedback Make You Not Try as Hard? A Study on Automotive Telematics

Vivek Choudhary
INSEAD, vivek.choudhary@insead.edu

Masha Shunko
Foster School of Business, mshunko@uw.edu

Serguei Netessine
The Wharton School, netessin@wharton.upenn.edu

Problem definition: Mobile and Internet-of-things (IoT) devices increasingly enable tracking of user behavior, and they often provide real-time or immediate feedback to consumers in an effort to improve their conduct. Growing adoption of such technologies leads to an important question: "Does immediate feedback provided to users improve their behavior?" We study immediate (close to real-time) feedback in the context of automotive telematics, which has been recognized as the most disruptive technology in the automotive insurance industry.
Academic/Practical relevance: Numerous automotive telematics providers claim unsubstantiated benefits from immediate feedback, while we still barely understand the implications of such feedback on user behavior. Given that feedback's effect sometimes is ambiguous, at the same time such feedback-providing devices' usage is increasing, it is important to study immediate feedback and identify the effect it has on human behavior, especially in important applications, such as automotive. This understanding is important given that other attempts to make driving safer have led to unintended consequences in the past.
Methodology: Using proprietary data on driving behavior, as measured by several parameters such as harsh braking, speeding, and steep acceleration, we investigate the impact of the driver's decision to review immediate feedback on driving behavior. We use instrumental variable regression to estimate the effect.
Results: Contrary to the claims from multiple telematics providers, we find that on average, users' driving performance after they review detailed feedback is nearly 14.9% worse than that of users who do not review their detailed feedback. This impairment in performance translates to 6.9%, or a one-year reduction in inter-accident time. Our results suggest that this deterioration is associated with increased speeding. Strong negative feedback (e.g., a sharp deterioration in performance) exerts a positive effect on short-term performance but this only happens for very large drops in performance (3% of cases). Furthermore, we demonstrate that drivers just below the insurance incentive thresholds exert greater effort following immediate feedback.
Managerial implications: Our results provide a key message to firms employing immediate feedback – specifically that such technology can yield unintended consequences. Furthermore, we show that drivers should receive only strong negative (but not positive) feedback to improve their performance. Finally, our results suggest that insurance incentives should be continuous, rather than a step function.

Keywords: Immediate Feedback; Empirical Operations Management; Behavioral Operations Management; Automotive Telematics

1. Introduction

Mobile and IoT devices increasingly allow tracking of user behavior that was not observable before, be it geolocation while running or heartbeat patterns during sleep. In addition to collecting and summarizing data, such applications and devices usually provide immediate feedback to users, typically in an effort to "improve" user behavior pertaining to a particular setting, e.g., to sleep better or run longer.

This paper focuses on telematics, an important IoT application in the automotive industry. In telematics applications, driving data are collected using a smartphone or an embedded device in a vehicle. Drivers receive feedback on their driving behavior after each trip, helping them analyze their driving skills, e.g., speeding habits, and potentially act upon this information. Given that driving is a major component of many operational systems, any small effect on driving behavior may carry massive implications for supply chains, pollution, accidents, and the economy in general. In view of driving's utility, it is no surprise that feedback on driving is coming to every part of operations for large fleets. For example, UPS provides daily feedback to its drivers (NPR 2014). Numerous automotive telematics providers claim unsubstantiated benefits from immediate feedback, while the implications of such feedback on user behavior are poorly understood. For instance, industry reports claim that "feedback from telematics devices can improve driving behavior for fleets" (Automotive-Fleet 2016). Companies like DriverMetrics work with logistics giants such as Greyhound and Kuehne + Nagel to provide feedback to drivers to change their behavior (DriverMetrics 2018). Furthermore, GreenRoad claims to help taxi and trucking companies achieve efficiencies (e.g., a reduction in maintenance costs and accidents) by providing real-time feedback to drivers (GreenRoad 2019). Based on our conversation with our industry collaborators, these claims are not substantiated with rigorous research and given the wide adoption and implications of telematics applications to vehicle operations, it is an important research question to investigate.

Telematics technology has become even more relevant now due to the emergence of the gig economy (e.g., ridesharing, crowdsourced deliveries, etc.), which has led to higher vehicle utilization (Cramer and Krueger 2016) and a growing number of drivers operating potentially unfamiliar vehicles in unfamiliar congested locations while increasingly relying on GPS-enabled devices. Such conditions create a potentially unsafe environment on roads, and fleet operators increasingly are using telematics to mitigate these concerns and better manage their drivers and fleets. Companies use data collected through telematics to lower insurance premiums and operational costs. For example, with the help of telematics, UPS has saved millions of dollars and gallons of fuel by implementing a "no left turn" policy (UPS 2016). Not surprisingly, telematics has been recognized as the most disruptive technology (Accenture 2018) in the $250 billion (2017) automotive insurance industry. Finally, telematics

applications have become ubiquitous as *smart city* initiatives are being launched worldwide. For example, Hasija et al. (2020) explain that, as urban living and congestion increase, it becomes imperative to "sustainably modify end-user behavior to ensure that social surplus is maximized…."

In this paper, we focus on the application of telematics in the automotive insurance industry, where insurance companies provide immediate driving performance feedback to users and incentivize individual drivers to drive better through premium discounts. This application is known broadly as usage-based insurance (UBI) and is expected to have 142 million subscribers worldwide by 2023 (IHS Markit 2016). While telematics devices already are providing feedback to drivers, we still barely understand the implications of such feedback on user behavior. Such an understanding is important, given that other attempts to make driving safer have led to unintended consequences. For example, mandating safety devices in automobiles has resulted in reckless driving behavior, causing more deaths and non-fatal accidents involving pedestrians (Peltzman 1975). Much of the existing literature suggests that "conventional" feedback (post-factum average performance, e.g., monthly utility bills) may lead to improved performance (Anseel et al. 2015). While these results are encouraging, our paper focuses on *immediate* feedback, which works differently from conventional feedback studied in extant literature. In the spectrum of feedback spanning from *conventional* to *real-time*, the timing of feedback in our application is very close to *real time*. The feedback is generated after each trip within seconds and it is sent to the driver through a push notification. Although a driver may not view it at all or right away, feedback is available immediately *after* completing the trip (in fact, a notification is provided about the overall trip score immediately after the trip's completion), and the driver can utilize this feedback *immediately* during the next trip. Very recent studies (Dahlinger and Ryder 2018, Gnewuch et al. 2018, Tiefenbeck et al. 2016) have shown that providing real-time feedback can be highly effective in some cases (e.g., utilities consumption, medicine intake). Given that conventional feedback's effect is sometimes ambiguous (Kluger and DeNisi 1996), it is important to study immediate feedback and its effect on human behavior, especially in important applications, such as automotive, which is what we focus on in this paper.

Using a novel dataset obtained through collaboration with Raxel Telematics, a Singapore-based company that creates customized telematics software/smartphone applications for fleet/insurance firms, we study the behavioral implications of telematics. Once a user opts in, the insurer tracks the user's driving behavior using sensors in the smartphone. In our case, users can review two types of scores after each trip: a trip score (the score for the latest trip) and a to-date score (an average score based on all the trips taken in the past). Users who obtain a to-date score above a threshold qualify for an insurance premium discount that increases in a step-wise manner with the score. A similar setup has been implemented by many other providers. For example, Rivigo, the largest technology-enabled fleet provider in India, offers a similar smartphone app for its drivers to review trip-wise driving

performance, and incentivizes them to improve. This performance is linked to drivers' 60% base pay (Rivigo 2019).

Our dataset comprises trip-level data collected via a smartphone application for 382 users, totalling 80,885 trips. After every trip, the application sends a notification to the users with their trip score, as well as an option to review detailed feedback via the app interface (we refer to trips completed after user-reviewed detailed feedback as *trips with detailed feedback*). Our application provides a rather unique setup that helps us study the impact of the decision to review immediate detailed feedback on end-consumers. First, we use novel trip-level data for diverse individuals receiving immediate feedback; they are not part of an organization or a group and have independent monetary incentives. Second, feedback is available in the application with no social cost, unlike in many organizational setups. Third, individuals receive the feedback without any information about other participants' driving behavior (no relative feedback). Fourth, users self-select to see their detailed feedback; therefore, the feedback-seeking behavior is intrinsic, rather than mandated. Fifth, in our setting, we know precisely whether/when people saw their detailed feedback. Given that we can track when users check their detailed feedback, our study is richer than many previous studies, in which we seldom know precisely whether/when users actually have read/observed the feedback.

Our results suggest that surprisingly, reviewing detailed feedback may have unintended consequences. On average, the driving performance of users who review the detailed feedback regarding their last driving trip is nearly 14.9% worse than the performance of users who do not review their detailed feedback, which is equivalent to a one-year reduction in inter-accident time. We attribute this effect to the overconfidence in driving abilities, which is a well-established phenomenon (e.g., Roy and Liersch 2013). Drivers who experience improvements in their scores become even more overconfident and take risks (namely, spend more time driving over the speed limit), leading to deterioration in performance, while drivers who experience very large decreases in their scores (which happens in about 3% of the trips with score decreases) improve during subsequent trips. On average, feedback leads to deterioration in performance, and namely, in the amount of time the subjects drive over the speed limit. Drivers also increase variation in their speeds (measured as the difference between the maximum and average speeds within a trip) after receiving feedback, which is associated in extant literature with a 2% increased probability of an accident (Quddus 2013). Although we show that feedback's overall effect is negative, we find that insurance-incentive thresholds are important, in that they make the impact of feedback consistently less negative.

This paper makes several contributions. We contribute to extant literature by examining lesser-understood behavioral impacts from immediate feedback on performance. First, we show that review of immediate feedback by drivers can have negative, rather than positive, effect on performance.

Second, we show that only very strong negative feedback benefits drivers, who improve after seeing it. Combining these findings with results from extant behavioral economics literature, we note that other behavioral interventions, such as "nudges," could be used in tandem with feedback to improve performance. For example, users, whose performance substantially has deteriorated, can be nudged to open their dashboards and study their feedback in detail. Finally, we demonstrate the critical role that incentive threshold plays in feedback effect: Drivers only react to it when they are just below the threshold. Thus, continuous discounts might be preferred over the step function.

2. Related Literature and Hypotheses Development

Our study builds on the rich body of work done in psychology, operations, and behavioral economics literature that seeks to understand a) the effect of feedback (conventional, immediate, or real-time) and b) the effect of economic incentives on user behavior and performance.

In the first literature stream, which studies the effect of feedback, very often, feedback has been associated positively with goal setting, self-awareness, and competency enhancements (Ashford et al. 2003, Renn 2001). Moreover, feedback provides information about prior performance and serves as a basis for evaluating one's ability to perform successfully on subsequent tasks (Bandura 1991). Due to these benefits, feedback systems have become ubiquitous in businesses and institutions. Given that human decision making often is biased (Kahneman and Tversky 1973), *feedback* can be used as a low-cost intervention to exploit human behavioral tendencies, such as *ahead-seeking* (utility gain from performing better relative to others), *behind-aversion* (utility loss from underperforming relative to others) (Roels and Su 2014), and/or *last place aversion* (Buell and Norton 2014). Our paper is an important step toward understanding feedback's effect in a complex operational setting related to, but different from, such environments as healthcare (Song et al. 2018), production (Schultz et al. 1998), and innovation (Mihm and Schlapp 2019).

Due to technological advancements, it has become possible to share automated feedback using smartphones, web browsers, or onboard electronic devices. Such feedback has found applications in industries ranging from utilities (Tiefenbeck et al. 2016) to transportation (Toledo and Lotan 2006), allowing for immediate feedback to improve performance. Very few papers have examined the effect of real-time feedback (close to immediate feedback), and those that have, show that it can be effective in improving behavior (Dahlinger et al. 2018, Gnewuch et al. 2018, Tiefenbeck et al. 2016). In our application, it is technologically possible to share real-time feedback, but it can be distracting and, therefore, dangerous, so the company instead defaults to immediate feedback. Immediate feedback often is provided automatically during lab experiments in operations management literature (such as in news-vendor games, e.g., Bolton and Katok 2008, Schweitzer and Cachon 2000, etc.). Unlike these studies, we focus on field data and study the *effect of* feedback, not decision biases in the *presence* of

immediate feedback. Moreover, unlike news-vendor experiments, driving involves a varying sequence of very different tasks without a clear "optimal" way to execute them.

In the second literature stream, financial incentives have been found to be instrumental in changing behavior and performance (Jenkins et al. 1998). This is important, as feedback and financial incentives have different functions and, if properly deployed, can exert a synergistic effect on behavior, as Stern (1999) concluded. The existence of an interaction effect between feedback and financial incentives, a conclusion this paper reached, is applicable in our study, as our drivers are incentivized for better driving through insurance discounts.

Using both feedback and financial incentives, feedback systems have been shown to be instrumental in improving operational performance. For example, a recent study (Soleymanian et al. 2019) using data from a UBI provider finds that individuals' overall scores improved by 9% and harsh braking decreased by 21% following enrollment in UBI. Contrary to our study, this paper studies the impact of being *in* the program (a one-time decision) and not the effect of *reviewing specific feedback* provided (which happens during the program). In addition, telematics implementations in fleets have shown significant improvements in driving safety (Leverson and Chiang 2014, Toledo et al. 2008) and fuel economy (Leverson and Chiang 2014). Similarly, many studies have focused on feedback systems' effect when employees are monitored during their day-to-day jobs. Managers use feedback for employee evaluation, while employees utilize feedback to improve their performance. For example, Blader et al. (2015) used a field experiment with a large U.S. transportation company with electronic on-board recorders that monitor drivers' performance automatically. Their results indicate that providing feedback leads to better performance. The feedback monitor is fitted on the trucks for drivers to monitor performance continuously, and alarms sound when a driver exceeds thresholds. Though real-time (close to immediate) feedback is a common theme in these papers, in all these cases, the feedback is forced on individuals, which is mostly feasible in fleet applications. In our case, drivers self-select to see their detailed feedback and do not receive any intermediate alarms about their driving behavior if they perform poorly during a trip for safety reasons. This is common in UBI applications in practice, such as Ingenie (Ingenie 2019) and MSIG UMax (MSIG 2018).

Finally, recent studies indicate a positive effect of real-time feedback in other settings. For example, Tiefenbeck et al. (2016) reported a 22% reduction in hot-water consumption when subjects are provided with real-time feedback on their water usage. Considering that in our case, drivers are provided with an explicit monetary incentive (discount), the drivers are motivated to improve their driving performance. As the app provides immediate feedback that allows for immediate action, we hypothesize that detailed feedback should lead to a higher performance score on the trip that immediately follows the observation of feedback.

It is important to note that feedback is only impactful if it can be linked to ways to improve. This is easily achievable when the task is single-dimensional (e.g., the user easily can interpret "your water usage is high" as a suggestion to use less water). However, in many operational systems (including when driving), providing feedback such as "your performance is poor" ("your driving score is low" in our application) would be insufficient, as it is not clear how the performance can be improved. Thus, detailed *feedback* helps users link improvement actions to performance (by *performance*, for purposes of this paper, we define it as short-term performance, which in our application is the trip score attained on the next trip). In what follows, when we refer to "reviewing feedback," we imply review of *detailed feedback* that includes specific information on various dimensions of performance (in our setting, this implies information that is available to the users inside the app, but not on a simple notification that only provides an aggregate score). Formally, we state our first hypothesis as follows:

H1a: Reviewing feedback exerts a positive impact on performance.

On the other hand, Kluger and DeNisi (1996) found in their meta-analysis that 38% of the studies report a negative effect of feedback on performance. They propose a theoretical framework (feedback intervention theory) to show that feedback can have a positive, negative, or insignificant effect. Concurring with their findings, recent studies have demonstrated feedback inefficacy on user performance (e.g., Volpp et al. 2017). Furthermore, several studies have demonstrated a negative effect of feedback. For example, studying accounting accuracy, Hannan et al. (2008) show that providing precise relative feedback can result in performance deterioration, as it induces ineffective strategies. Rolim et al. (2017) observe that bus drivers do worse after real-time feedback because they focus on safety parameters, while overlooking others.

As most people believe that they drive better than the average driver (Roy and Liersch 2013), it is plausible that feedback in our setting can be disconfirming (i.e., lower than our own self-assessment), resulting in poor performance due to threat to self-concept (i.e., the threat to one's belief about his/her abilities, Green et al. 2017). Although drivers in our study do not feel peer pressure, a common scale of scoring (i.e., every driver is scored on the same 0-100 scale) can induce a comparison effect that can lead to performance deterioration (Ashton 1990). Therefore, we hypothesize that feedback in the case of telematics may exert a negative effect:

H1b: Reviewing feedback exerts a negative impact on performance.

As is evident from the above discussion, due to conflicting results from previous studies, the effect of feedback is contextual, and we do not have a strong prior on the direction of the effect of feedback on driving performance, although telematics providers claim positive effects of immediate feedback. Next, we suggest a possible mechanism behind this seemingly ambiguous effect of feedback by arguing that based on the recent trend in users' performance, feedback may sound either positive or negative to

the user, e.g., if a driver's score recently has increased, the feedback will carry good (positive) news to the driver, while if the driver's score recently has decreased, the feedback will bring bad (negative) news to the driver. As a result, we theorize that drivers will react differently to feedback based on whether the feedback is perceived as negative or positive. Prior literature again proposes different mechanisms that can lead to either positive or negative reactions to good or bad news. Thus, we again developed competing hypotheses.

In considering negative feedback, several papers proposed mechanisms by which negative feedback leads to further performance deterioration due to negative updates of users' beliefs and/or due to demoralization. For instance, using a field experiment with secondary school students, Fischer and Wagner (2019) showed that feedback regarding a deterioration trend in performance given to students immediately before an exam hindered subsequent performance by negatively updating their beliefs. Similarly, Azmat et al. (2018) find that students who received feedback exhibited performance deterioration because they underestimated their abilities. Barankay (2012) experimented with salespeople and found that providing private feedback reduced sales performance by 11% because of the demoralization effect: Employees reduced their efforts when they received lower-than-expected performance feedback. Similarly, Gjedrem (2018) found that subjects who report poorer ability hindered their performance substantially when feedback was provided. In our case, it is possible that negative feedback may lead to performance degradation due to attributing performance deterioration to external factors or chance because of attribution bias (Billett and Qian 2008). For example, Grossman and Owens (2012) found that in the presence of overconfidence, which is typical in our setting, negative feedback can be attributed to *unlucky* circumstances. In our case, it is possible that small reductions in scores are being attributed to circumstances. Such decreases can be viewed as an outcome of chance or uncontrollable factors that are not affected by drivers' effort; therefore, drivers can reduce their efforts in anticipation that it will exert no effect on subsequent performance. Using parallel arguments to the above logic, positive feedback can lead to an increase in morale and positive update of drivers' beliefs, resulting in improved performance. These arguments lead us to propose H2a:

H2a: A review of positive and negative feedback leads to a differential impact on performance:

H2a.1: Positive feedback (improvement in recent performance) leads to improvement in subsequent performance.

H2a.2: Negative feedback (deterioration in recent performance) leads to deterioration in subsequent performance.

On the other hand, several studies find that negative feedback leads to improved subsequent performance. For instance, Ashford and Tsui (1991) found that negative feedback (associated with seeking critical appraisals that might hurt) leads to better performance. In a setup similar to ours,

Soleymanian et al. (2019) show that providing daily feedback on negative changes (or an increase in harsh braking) in the past two trips leads to better performance in the same parameter the next day. Furthermore, Burgers et al. (2015) illustrated a positive effect of negative feedback on performance with temporal nuance in a study employing computer games. They show that negative feedback (e.g., "you perform poorly") exerts a positive impact on performance in the short term. One mechanism that explains the positive effect of bad news is that negative feedback can shatter users' self-confidence, which is known to decrease optimism bias and can lead to drivers exerting more effort to improve (Moores and Chang 2009).

However, after observing an increase in their scores relative to the previous trip, which comprises positive feedback, drivers' self-efficacy (or confidence) increases. An increase in confidence in one's own driving ability leads to an increase in optimism bias (Krueger and Dickson 1994), which has been found to lead to risk-seeking behavior (Sharot 2011). In our context, this may imply speeding and harsh braking, which leads to further reductions in driving performance. Therefore, we create a competing hypothesis, H2b, regarding the direction of positive and negative feedback's impact on performance, and formally, we state our hypothesis as follows:

H2b: A review of positive and negative feedback leads to a differential impact on performance:

H2b.1: Positive feedback (improvement in recent performance) leads to deterioration in subsequent performance.

H2b.2: Negative feedback (deterioration in recent performance) leads to improvement in subsequent performance.

Figure 1: Differential effect of feedback on performance

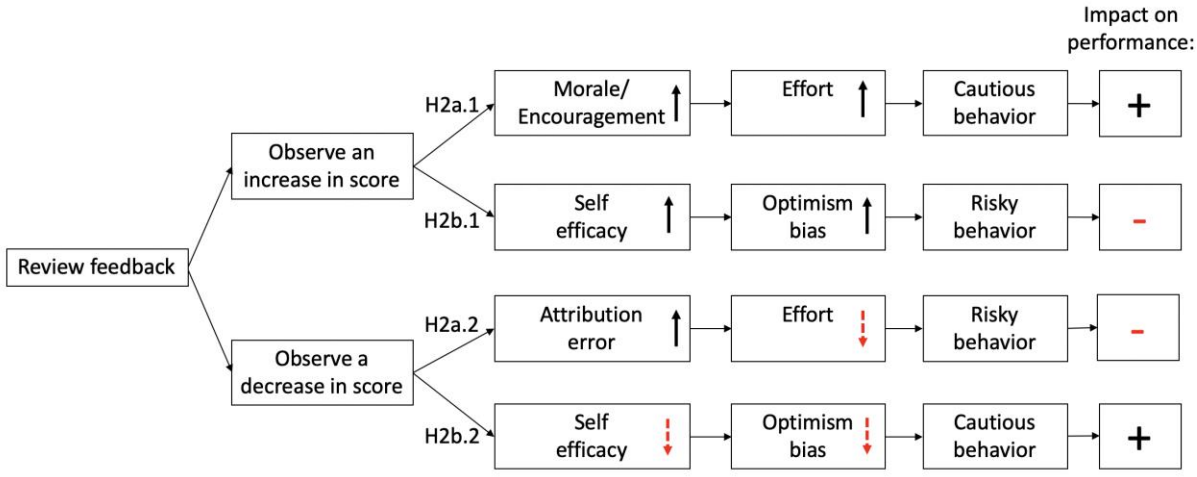


Figure 1 illustrates the competing effects of feedback that we propose. Understanding which hypothesis is supported in our case will suggest which behavioral mechanisms might determine drivers' reactions to feedback.

'Goal-gradient' theory suggests that people whose performance is close to the goal tend to exert higher effort to improve performance, and this effect is especially pronounced for people whose performance is closer to the first reward (Kivetz et al. 2006). After reaching the reward threshold, people tend to reset their efforts. For example, in reward programs, people tend to spend more effort to gain the next status (e.g., airline loyalty programs). Using lab experiments, it has been found that people pay more attention to a goal as they get closer to attaining it, and that visualization (e.g., through progress bars) increases commitment and effort (Cheema and Bagchi 2011). This effect of increasing effort near the goal, termed the "goal looms larger" effect, is well-established in extant literature (Förster et al. 1998). Similarly, Berger and Pope (2011) show that being slightly behind increases effort. Analyzing basketball games, they find that players exert more effort when slightly behind at half-time. As a result, such teams win more games than teams that lead prior to half-time. Furthermore, using a lab experiment on individuals, they conclude that in individual tasks, fixed standards (such as incentive thresholds) may act as reference points, and that being slightly behind these reference points encourages effort. In our case, we predict that drivers who look at their feedback and observe that they are closer to achieving the next threshold will exert more effort to improve performance. Therefore, observing closeness to the next incentive threshold will moderate the effect of feedback on drivers' performance positively. Once the goal is reached, the motivation to do well will diminish (Lee et al. 1997), which is consistent with diminishing economic returns from the effort. Therefore, we hypothesize that proximity to the threshold from below will exert a positive effect on performance:

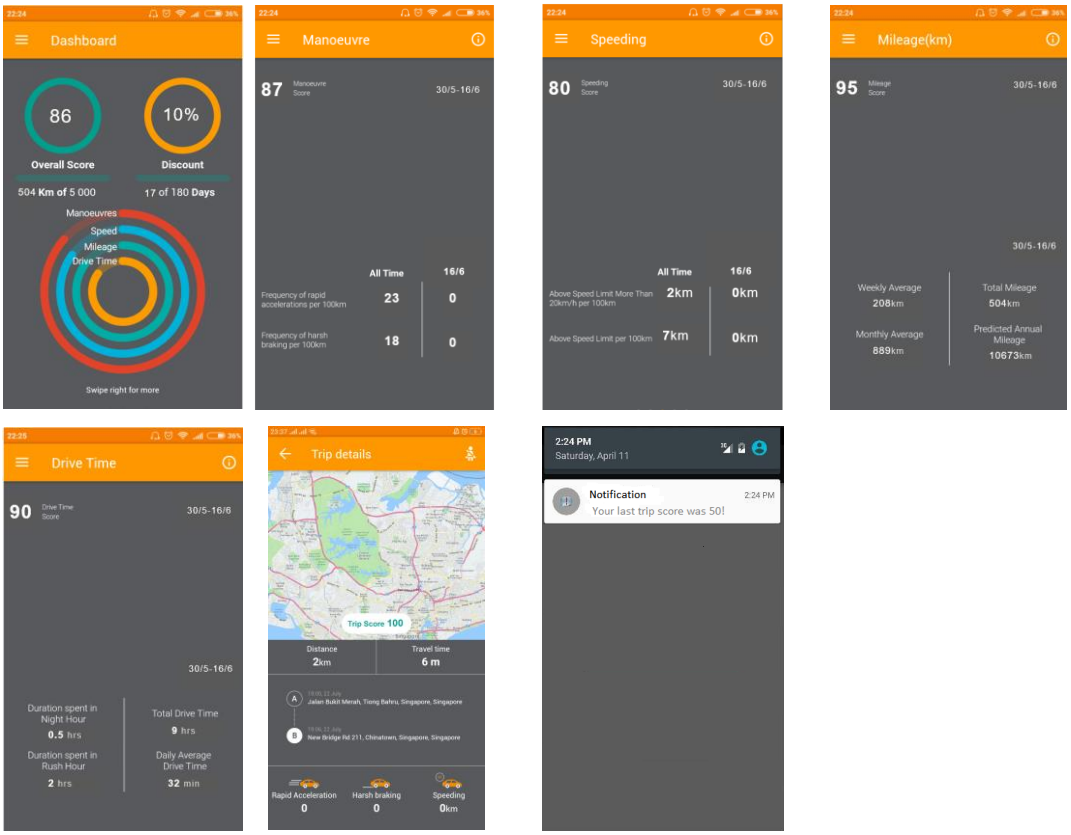
H3: Approaching the next incentive threshold positively moderates feedback's impact on performance.

3. Data and Analysis

3.1. Data and Variables

Our dataset comprises trip-level records for private car drivers enrolled in UBI offered by a large insurer. The users enroll in the program using a smartphone application (app). For each trip, the app collects start and end time stamps, origin, destination, GPS coordinates of the entire trip, and miles driven. In addition, the app captures driving behavior on three dimensions—speeding, acceleration, and braking—and assigns a trip score on a scale from 0 to 100 based on a proprietary algorithm that combines the assessments on these dimensions using a variety of sensors built into the smartphone (such as a gyroscope, GPS, and accelerometer).

Figure 2: Smartphone Application (Screenshots and Notification)



The app comprises one dashboard and five other screens (see Figure 2 with screenshots). Along with the dashboard, a separate screen is provided to review individual trip scores (second screen in Figure 2). The first screen is the dashboard that provides a summary of the to-date performance along various dimensions of driving behavior. In the lower part of the dashboard, the maneuvers ring measures acceleration and deceleration behavior, and the speed ring captures speeding behavior. The dashboard also provides the discount percentage that a user can obtain with the current to-date performance at the time of policy renewal. The screen provides graphical scores regarding how a user performs across maneuvers (i.e., accelerations and braking behavior), mileage (projected kilometers one would drive in a year), speeding, and drive time (driving during rush hours or during nighttime hours) graphically. A full solid ring denotes a score of 100. Partial solid rings in Figure 2 denote respective scores out of 100. The app also provides a mileage score based on the estimated kilometers a user would drive in a year. A higher number of kilometers driven is associated with a higher probability of accidents; therefore, higher mileage is related to lower scores. In our app, users are penalized for mileage if they drive more than 10,000 km a year. Users can swipe right to the other screens in the app to see additional details of their driving performance, e.g., the number of steep accelerations they have made per 100 km travelled, as well as trip-level detailed performance. Similarly, detailed analysis regarding steep acceleration,

speeding, and harsh braking is also available to users. The difference between the detailed feedback screen and notifications should be noted. The last screen in Figure 2 shows the notification sent to every user after each trip (provided the user enabled push notifications); it shows only the trip score, while detailed feedback includes performance on past trips across various variables (acceleration, braking, speeding, etc.). The detailed feedback screen also shows to-date performance (net score that leads to a discount) and expected discounts. Moreover, from the detailed feedback dashboard, users can visit supplementary screens with more details about particular dimensions of their scores. Our correlational analysis shows that whenever a driver looks at the detailed feedback dashboard, he or she also visits all other screens (correlation > 0.99), suggesting that drivers care not only about the net score (which determines their monetary incentive), but also about the detailed information, which may speak to their intrinsic motivation to improve.

The app uses a state-of-the-art machine-learning algorithm to create unique driving signatures in the first few trips. This signature helps identify whether a driver is behind the wheel or is a passenger in the car. Furthermore, the GPS signal complements the smartphone application throughout the trip to identify the mode of transport, even walking. These algorithms are used routinely in the industry.

Evidently, three self-selections happen during the UBI process. First, drivers decide to insure with the focal insurance company, probably because the company offers them a competitively priced policy. Second, the drivers opt into the optional UBI program. Third, drivers choose to seek feedback by visiting the detailed trip-evaluation screens. Unfortunately, our data do not allow us to study or address the first and second self-selection issues. Instead, we focus in this paper on the effect of feedback on the driving behavior of users who choose to enroll in the UBI and receive immediate feedback on their driving behavior. This population is, in fact, most relevant to insurance companies because they have neither the means, nor the desire, to force the app on all insured, and they have no way to insure everyone. Naturally, while the effects that we study have a relatively short time horizon, it is possible that, in the very long term, they can begin to affect who the company insures and who decides to enroll in UBI.

Any driver can register for the telematics application that the insurer provides, as there is no cost or risk to sign up. A good score can qualify a user for a discount, but a bad score currently does not increase the premium (i.e., there is no penalty). Once drivers sign up, they need to keep the application running for a minimum of six months and drive at least 5,000 km. After successful completion of these conditions, the aggregated to-date score is computed and frozen, and drivers can obtain a discount on the insurance policy renewal fee based on the final to-date score: no discount for scores below 70; 5% for scores between 70 and 80; 10% for scores between 80 and 90; 15% for scores between 90 and 95; and 20% for scores above 95. The discount on the premium provided is on top of the no-claims discount.

We have 81,839 recorded trips in the original dataset. To ensure that our trips are comparable, we dropped 239 trips that originated from or were destined to Malaysia (neighboring country). We further removed trips with lengths below 0.3 km and above 70 km (excluding another 488 observations). Similarly, we excluded users with fewer than 10 trips in total (227 trips). Our final dataset comprised 80,885 observations for 382 users over a period of 140 days. This was the longest time window we obtained in which we have a stable population of drivers, as well as an unchanged app. Next, we describe the variables used in our analysis, in which *i* denotes users, *j* denotes trips, and *t* denotes date when a trip occurred:

$oscore_{ijt}$. Our outcome variable, which we refer to as (overall) trip score. After each trip, the app sends a notification to the users' smartphones with their overall scores for the previous trip on a scale of 0 to 100. This score is derived according to a proprietary algorithm based on three user behaviors (speeding, acceleration, and braking) and two trip aspects (hour during which the trip was taken and distance travelled).

$todatescore_{ijt}$. This score reflects to-date performance of the user *i* till the trip *j* that occurred on date *t* and is used by the insurer to provide a discount after the program's completion. This score also is calculated using a proprietary algorithm from the app provider.

While each user may observe their score on a push notification (provided the user has enabled notifications in the app settings), some users choose to open the dashboard and review their trip score, to-date score, potential incentives, and performance on various driving dimensions. We call this event a *detailed feedback review*, and we know when users open the app and which screens they visit as the app captures this event. While push notifications (e.g., "Your last trip score was 50!") can be viewed as some sort of feedback, they do not provide any details on what causes the score to be low/high, nor any information on how to improve performance. Thus, we assume that viewing detailed feedback allows the driver to better understand how to improve or to understand where to slack off. Moreover, some users will not even enable push notifications from the app, and even if they do, it is not clear whether drivers actually notice the push notifications. Later in the paper (see Section 3.3a), we conduct a robustness test to verify that drivers do not change behavior based on push notifications, which supports our operationalization of "feedback" as a detailed feedback review.

Thus, we define the binary variable *feedback:*

$feedback_{ijt}$ = 1 if a user *i* visits the main dashboard before the current trip *j*, which takes place at time *t*; 0 otherwise.

$perdailyhr_{ijt}$, $perrushhr_{ijt}$, $pernighthr_{ijt}$. Drivers' scores are penalized if travel happens during rush hours or at night, times associated with a higher probability of accidents. To account for these variables' effect on scores, we control for the trip-duration percentage that occurs during night and rush hours.

For each trip, we divide the minutes travelled during daily/night/rush hours by the total trip duration. Figure 3 illustrates the classification of hours for different days as set by the insurer. Note that one trip can cover multiple types of hours, e.g., the first few minutes can be driven during a rush hour, followed by a few minutes travelled during a regular daily hour.

$freq_{ijt}$. Users who drive rarely may behave systematically differently than the frequent drivers, e.g., frequent drivers have more experience driving than non-frequent drivers. Although drivers are required to drive a minimum of 5,000 km during 180 days to participate in the study, some drove less during the window of our data sample (142 days). To control for this effect, we calculate the number of trips completed by user *i* since enrolment in the program prior to trip *j* that occurred on day *t* divided by the number of days since enrolment in the program by day *t*.

$fbfreq_{ijt}$. Users differ in how often they seek feedback, and it has been shown that feedback frequency affects performance (Lurie and Swaminathan 2009b). For user *i*, we calculate what percentage of trips taken before trip *j* that occurred on day *t* followed a detailed feedback review.

$perphoneuse_{ijt.}$ Through this app, we observe whether the phone was used during a trip and calculate the percentage of trip time during which the phone was used. Phone usage usually is distracting and potentially can affect the overall score. For this reason, governments have started putting regulations in place, e.g., phone usage for calls and texting is prohibited while driving in Singapore (Singapore Road Traffic Act 65B 2015). Beyond regulations, road transport authorities are investing in behavioral interventions to curb distracted driving (PennMedNews 2018). Unfortunately, we cannot be sure of the reasons for phone usage (e.g., navigation, calls, or texting). Removing this variable does not alter our results.

Figure 3: Definition of Rush Hours, Nighttime Hours, and Daily Hours

Hour of the day	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Working days	N						D		R			D							R			D		
Non-working days	N						D																	

N - night hours, D - daily hours, R-rush hours

In addition to the above variables, we use two lagged variables in our analysis. We code these variables $ltodatescore_{ijt}$, which is the to-date score of user *i* one trip before trip *j* and $loscore_{ijt}$, which is the $oscore_{ijt}$ of user *i* one trip before trip *j*. The latest to-date score observed by the user on the app, as well as the previous trip's score, can influence the user's driving behavior on the subsequent trip. Thus, we use these variables as controls in our econometric model.

We provide the descriptive statistics for all the variables in the first four columns in Table 1. On average, trips have the highest percentage of day hours, with a mean of 62.1%, followed by rush hours, with a mean of 34.9%. The users in our sample take an average of 3.7 trips per day. Due to strict

regulations, it is unlikely that our dataset contains any taxis. Nearly one third of the trips followed a detailed feedback review (*feedback* has a mean of 0.33). It is evident from the table that our outcome variable *oscore (*for which statistics are similar to the ones of *loscore)* is right-skewed with a mean of 81.39 on a scale of 0 to 100. 52,412 trips have happened without any phone usage during the trip, while the average phone usage was 2.47% of the time during a trip.

Table 1 also demonstrates the correlation structure in our data (Columns 1 to 9). For our analysis, we keep only *perdailyhr* and *pernighthr*. We do not report any other collinearity issues in the variables used. *Feedback* is a binary variable; therefore, we report *polyserial* correlations.

We start analyzing our data by conducting a *t*-test to observe any differences between the mean scores for the control trips (for which users did not review their feedback) and the treatment trips (for which users reviewed detailed feedback). Out of the 80,885 trips, we have 26,313 treatment trips and 54,572 control trips. We find that the mean score for the treatment trips (80.4) is 1.9% lower than that of the control trips (81.9). Thus, we see that trips with detailed feedback have lower scores, on average, compared with the trips without detailed feedback. Furthermore, we conducted the Kolmogorov-Smirnov equality-of-distributions test to compare the distributions of the trips with detailed feedback to the trips without detailed feedback. Our results suggest that the two distributions are significantly different, with $p<0.01$.

Table 1: Summary Statistics and Correlation Structure of the Variables

	Summary statistics												
	Mean	SD	Min	Max	(1)$^{+}$	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1) feedback	0.33	0.47	0.00	1.00	1.00								
(2) perrushhr	34.99	46.74	0.00	100.00	-0.09*	1.00							
(3) perdailyhr	62.10	47.44	0.00	100.00	0.08*	-0.94*	1.00						
(4) pernighthr	2.91	16.41	0.00	100.00	-0.06*	-0.13*	-0.22*	1.00					
(5) freq	3.66	1.43	0.12	13.00	-0.13*	-0.02*	0.02*	0.01*	1.00				
(6) fbfreq	35.20	19.06	0.00	100.00	0.53*	0.02*	-0.02*	-0.00	-0.23*	1.00			
(7) perphoneuse	2.47	6.87	0.00	100.00	0.02*	-0.02*	0.01*	0.02*	0.02*	0.03*	1.00		
(8) ltodatescore	72.59	16.64	22.00	100.00	0.04*	0.01	0.00	-0.03*	-0.50*	0.09*	-0.03*	1.00	
(9) loscore	81.39	16.81	22.00	100.00	-0.05*	-0.12*	0.14*	-0.06*	-0.01*	0.01*	-0.03*	0.20*	1.00

**p<0.05, +polyserial correlation, oscore has similar summary statistics as loscore.*

3.2. Econometric Model

To test our hypotheses, we formulated a linear regression model to evaluate the impact of reviewing detailed feedback prior to the focal trip on driving behavior during the focal trip. Since decisions to review feedback are likely to be endogenous, we use instrumental variable regression for our main analysis. Our data set is an unbalanced panel, as users in the sample complete different numbers of trips during the study period. We specify the following econometric model:

$$oscore_{ijt} = \alpha_0 + \beta feedback_{ijt} + \gamma Controls_{ijt} + date_t + individual_i + orig_j + des_j + \varepsilon_{ijt.} \quad (1)$$

We include a $Controls_{ijt}$ vector to account for the potential confounding factors summarized in Table 1. To account for the unobserved heterogeneity among dates, such as weather impact, we use *date* dummies. To control for time-invariant individual unobservable characteristics, such as gender and education, we use *individual* fixed effects. As our data captures about five months of observations, we believe that individual fixed effects also will take care of unobserved driving skills and vehicle types that individuals use. To control for location-specific characteristics—such as commercial office area, number of traffic lights, and road width—we include origin and destination dummies ($orig_j$ and des_j). To obtain them, we map each address to the corresponding district area using the standard location classification established by the Singapore government. Our origins and destinations cover Singapore's 64 district areas. We have repeated observations for each individual, so we cluster the standard errors at an individual level to address potential correlations among the driving scores. We note that our $Controls_{ijt}$ vector includes a lagged dependent variable $loscore_{ijt}$, which, in combination with fixed effects, can lead to well-known Nickell bias (Nickell 1981) in a dynamic data panel. We therefore try omitting this control and our results remain essentially the same.

As a preliminary analysis, we run OLS regression for model (1) and report the results in the Appendix, Section E.1. As we do not assign feedback to trips randomly, some other underlying characteristics of the trips may influence the scores, which may be attributed mistakenly to feedback. To account for this non-random assignment, in Section E.2 in the Appendix, we employ matching methods (propensity score and nearest-neighbor matching). As a summary, our estimated effect size using OLS is within the range of the results obtained using matching.

3.2.1. Instrumental Variable Analysis

To test Hypothesis 1 robustly, we must recognize that looking at detailed feedback likely is an endogenous decision. We realistically can envision a mechanism such that some of the unobserved variables may affect both *feedback* and *oscore*. Such omitted variables will lead to an underestimation of the effect of feedback on performance and, potentially, a biased estimation. We conduct a Hausman test, and we cannot reject the null hypothesis that *feedback* is an endogenous variable. Therefore, to address the endogeneity of the decision to review detailed feedback, we use an instrumental variable approach.

As an instrumental variable, we propose the percentage of time that the mobile network was disrupted between trips (outage), which is an exogenous shock affecting users' ability to receive and review feedback, but not the ability to drive, as we do not consider outages during trips that may disrupt navigation. To construct this variable, we use data from http://downdetector.sg, which publishes user-

reported network disruptions throughout the year for three telecom providers (Singtel, Starhub, and M1) in Singapore (these providers comprise nearly 100% of the mobile users in the country). The website checks and validates this data from a series of sources to ascertain network disruptions. The website does not provide a precise timestamp of outage resolution, so to estimate outage lengths, we look at the distribution of outage reports over time: The number of reports increases following the first report of an outage, then drops sharply after the resolution. We designate the time between the first report and the moment of sharp drop in reports as the duration of an outage. We then use $outage_{ijt}$ to denote the percentage of time the outage lasted during the time interval between the end of trip *j-1* that occurred on date t and the moment of time the user *i* looked at feedback after the end of the trip. If the user did not look at feedback between trips *j-1* and *j*, we consider the time interval between the end of trip *j-1* and the beginning of trip *j*. The average time between trips is 453.9 minutes, while the average time between the end of trip *j-1* and feedback view is 213.3 minutes (see details in Table 2).

Table 2: Timing of trips and feedback

Variable	N	Mean	SD.	10%	50%	90%
Time between trips (min)	80,503	453. 9	1421.4	19.9	161.0	928.8
Time from trip's end to feedback view (min)	26,313	213.3	376.7	22.5	179.3	310.9

Note: % denotes respective percentiles

We believe that the users have less opportunity to check their feedback if the outage lasted for a large percentage of time between trips and, thus, *outage* should elicit a direct effect on the *feedback* variable to satisfy the relevance condition. Out of all 80,885 trips, 16,270 are affected by an outage event. The outage variable varies from 0 to 100%, with an average of 2.23%. The 99th percentile value of outage variable is 47.04%.

To check whether the exclusion condition also is satisfied (*outage* does not impact *oscore* directly), we explore several potential mechanisms. First, we check and confirm from the company websites, as well as from news wires, that the outages are technical faults and are not due to poor weather that may result in excessive traffic that, in turn, can exert a direct impact on *oscore*. Next, we check for an occurrence of large-scale events (e.g., an F1 race) during our observation window, which could be an omitted variable causing increased usage of phones and difficult driving conditions (due to traffic diversion) that could affect *oscore* directly. We did not identify any such events during the study period. Finally, one could argue that, immediately following a network outage, users may use the phone more, and if this increased usage spills over to the next trip, the *oscore* may be impacted directly. To satisfy the exclusion condition with this mechanism, we control for phone usage in the model, such that the *outage* is not correlated with the error term in the *oscore* regression *conditionally* on this control.

We use the method provided in Wooldridge (2002) to estimate the effect of feedback on performance, as our treatment variable is binary in nature (for a practical application of this method, see Adams et al. 2009). This method is advantageous to traditional 2SLS as it obtains correct standard errors (Wooldridge, 2002). Namely, in this method, we first predict probabilities of users looking at their detailed feedback by estimating a *probit* model. Then, using the predicted probabilities as instruments, we run a 2SLS model with *feedback* as an endogenous variable. Specifically, the three-step method is: 1) We estimate a binary response model of $feedback$ on *outage* (our instrumental variable) and control **x** to estimate the fitted probabilities of $\widehat{feedback}$. 2) We then regress *feedback* on fitted $\widehat{feedback}$ and control **x** to estimate the fitted values of $\widehat{\widehat{feedback}}$. 3) Finally, we regress $oscore$ on $\widehat{\widehat{feedback}}$ and control **x**. In the first step, we estimate a *probit* model using *outage* and other control covariates as follows:

$$\Pr(feedback = 1|\mathbf{E}) = \Phi(\gamma\mathbf{E}), \quad (2)$$

in which **E ={***outage***, x}** is the set of exogenous variables. The results of the first-step estimation are reported in Table 3. To estimate the parameters and ensure convergence, we had to drop nine observations (outliers) in which the times between two trips were more than seven days. In Step 1, we obtain a significant negative coefficient for *outage* (-0.025, p<0.01) as reported in Table 3. The negative sign indicates that, as predicted, higher *outage* is associated with a lower probability of users reviewing their dashboards. Next, using the predicted probabilities, we run a 2SLS model to estimate the effect of *feedback* on *oscore* (full results of the first stage with all controls are reported in Table E.7 in the e-companion).

Table 3: First Step: Instrumental Variable Regression (Probit Model)

VARIABLES	*feedback*
outage	-0.025***
	(0.003)
Constant	-3.693***
	(0.741)
Controls	yes
Observations	79,639

*Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1, estimation done with individual, date, origin, and destination fixed effects and all controls (perrushhr, perdailyhr, pernighthr, freq, fbfreq, perphoneuse, ltodatescore, and loscore). SEs clustered at individual level.*

During the first stage of 2SLS, we obtain an F-stat >10, indicating that our instrument is strong (Staiger and Stock 1997). Further, by conducting Anderson-Rubin test (χ^2=36.69, p<0.01), and under-identification test (significant at p<0.01), we confirm the instrument's sufficiency. We report the results from the three-step estimation in Table 4, Model 1. We find a negative and significant coefficient (-2.508, p<0.01). On average, users who review their detailed feedback prior to a trip perform worse on

the upcoming trip by 2.508 points, or approximately 14.9% in terms of the standardized effect size (measured as a multiple of standard deviation of *oscore*), than the users who do not review their feedback. Relative to the results reported in Table E.1, the effect size is nearly 2.5 times higher than when estimated without the instrumental variable.

Furthermore, as a robustness test, we also estimate a model assuming that *feedback* is a linear, rather than a binary, variable with 2SLS estimator, and we report consistent results in Model 2 of Table 4. To summarize, our instrumental variable analysis is consistent with simple OLS regression and with matching methods in supporting H1b and rejecting H1a. However, the size of the effect nearly triples once we utilize instrumental variables: Both OLS and matching methods underestimate the impact of feedback on performance.

Table 4: Results of Instrumental Variable Regression

	(1)	(2)
VARIABLES	*oscore*	
feedback	-2.508***	-2.937***
	(0.382)	(0.448)
perdailyhr	0.146***	0.146***
	(0.003)	(0.003)
pernighthr	-0.149***	-0.150***
	(0.005)	(0.005)
freq	0.189	0.169
	(0.131)	(0.127)
fbfreq	0.038***	0.043***
	(0.011)	(0.012)
perphoneuse	0.033**	0.034**
	(0.013)	(0.013)
ltodatescore	0.004	0.005
	(0.034)	(0.033)
loscore	0.072***	0.072***
	(0.005)	(0.005)
Constant	78.723***	78.635***
	(10.517)	(10.477)
R-squared	0.393	0.392
Model	3-steps	2SLS

*Robust SEs in parentheses *** p<0.01, ** p<0.05, * p<0.1, estimation done with individual, date, origin, and destination fixed effects and all controls. SEs clustered at individual level. N = 80,503.*

After discussing the results with the collaborating firm, we can estimate the economic size of the effect. Based on historic data, the firm can link the scores to the number of years until the next accident. For an average driver in our data set with a mean score of 81.4 (average inter-accident time of 14.6 years), one score-point reduction in performance will reduce the time until the next accident by 0.4 years, everything else being equal (i.e., all control variables remain the same, e.g., drivers traveling between similar locations, on the same day, etc.), thereby significantly increasing the probability of having an accident. Therefore, our results indicate that *feedback* reduces the inter-accident time by 6.9%

(=2.508×0.4/14.6%) or approximately by one year (=0.4×2.508).

So far in testing H1, we used the overall score as a dependent variable. To better understand the reasons for our key results, here we analyze the effect of *feedback* on different dimensions of driving, such as harsh braking, sharp acceleration, and speeding. This is particularly of interest, as the scoring algorithm is proprietary and, therefore, we cannot precisely identify the effect of feedback on specific dimensions of driving behaviour from the overall result. *acclr_rate* and *brk_rate* measure the number of incidences of sharp acceleration and harsh braking per 100 minutes of a trip. *accl_rate* has a mean of 1.68 (with a standard deviation of 5.64) while *brk_rate* is 2.32 on average (with a standard deviation of 6.73). Our third variable, *overspeedrate*, denotes distance (in km) travelled while speeding per 100 minutes of a trip which is on average 2.12 (with a standard deviation of 5.08). We report the results in Table 5. The first three models suggest that *feedback* deteriorates driving behavior by increasing distance travelled while speeding, but it exerts no effect on harsh braking or sharp acceleration.

Table 5: Effect of Feedback on Various Dimensions of Driving Behavior

	(1)	(2)	(3)	(4)
VARIABLES	*acclr_rate*	*brk_rate*	*overspeedrate*	*max-avg speed*
feedback	0.147	-0.017	0.798***	2.064***
	(0.108)	(0.116)	(0.124)	(0.297)
Constant	1.188	1.401	-5.374***	29.939**
	(1.080)	(1.330)	(1.653)	(14.523)
Controls	yes	yes	yes	yes
R-squared	0.086	0.093	0.160	0.160

*Robust SEs in parentheses, *** p<0.01, ** p<0.05, * p<0.1, with individual, date, origin, and destination fixed effects and all controls (perrushhr, perdailyhr, pernighthr, freq, fbfreq, perphoneuse, ltodatescore, and loscore). N = 80,503. SEs clustered for individuals, estimation with instrument.*

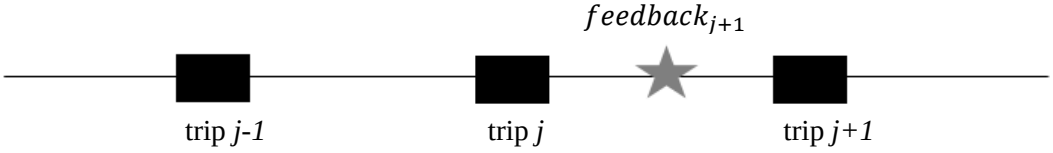
Nicholas and Gadirau (1988) found that "accident rates do not necessarily increase with higher average speed but do increase with higher speed variance." Thus, we also look at how feedback impacts variation in speed. In particular, we measure variation in speed per user as the difference between the maximum and average speed; this metric has been found to be highly correlated with accidents (Quddus 2013, Tanishita and van Wee 2017). We present the analysis next, and we find that the feedback view has a positive association with higher speed variation. Namely, our results suggest that feedback increases speed variation by 2.064 (Column 4, Table 5). For an average user with average variation of 31.74 km/hr (*maxspeed-averagespeed*), this effect is equivalent to a 6.50% (2.064/31.74%) increase in variability. Quddus (2013) suggests that it will increase the probability of an accident by nearly 2% on average (a 1% variation increase is associated with 0.3% increased probability of accident).

3.2.2. Positive vs. Negative Feedback

To test Hypotheses 2a and 2b, we first need to operationalize positive and negative feedback.

Consider three trips as illustrated in Figure 4. We omit subscripts *i* and *t* for ease of exposition. Our focal trip is *j+1*. Drivers can observe either an increase (a gain) in their scores or a decrease (a loss) in their scores in the past two trips from *j-1* to next trip *j* before viewing feedback.

Figure 4: Definition of Gain and Loss



For feedback observed prior to trip *j+1*, we define *gain* or *loss* based on the change in trip scores from trip *j-1* to trip *j*. Let $delta_{j+1} = oscore_j - oscore_{j-1}$, then:

$loss_{j+1} = |delta_{j+1}|$ if $delta_{j+1} < 0$, and $loss_{j+1} = 0$ otherwise,

$gain_{j+1} = delta_{j+1}$ if $delta_{j+1} > 0$, and $gain_{j+1} = 0$ otherwise.

Note that both *gain* and *loss* are defined using strict inequalities, such that these variables capture the change in the coefficient relative to the case of no change in performance score. Also, for clear interpretation of results, we have taken the absolute value for loss. Next, we estimate the effect of *feedback* on trip score accounting for whether the driver observed *a gain* or *a loss*. Notice that the user can only observe *a loss (gain)* if he or she actually incurred *a loss* (*gain*) before looking at feedback.

As is evident from Table 6, the *gain*×*feedback* term has a negative significant coefficient (-0.059, p<0.01) for the total effect of positive feedback of -1.972-0.059**gain*, which indicates that upon reviewing an increase in score, drivers perform worse during the next trip, which supports Hypothesis H2b.1, but not H2a.1. This suggests that the driver's reaction to reviewing positive feedback might be due to an *increase in optimism bias*, rather than due to an increase in morale or positive belief update, which would lead to an increase in score. This finding is consistent with the known notion that drivers often are overconfident (Roy and Liersch 2013), and as they already have high confidence in their driving ability, positive feedback only reassures them in their ability to engage safely in risky driving behaviors. Also note that the main effect of *gain* is insignificant, indicating that it is not the performance increase itself that leads to worse performance during the next trip, but the effect of *letting the drivers know that their performance increases* leading to decreases in the next trips' scores.

On the other hand, when the drivers' performance deteriorates, the effect changes: The *loss*×*feedback* term has a positive significant coefficient (0.041, *p<0.05*). Given that the subject incurred a loss, the net effect of feedback after observing a decrease in score is less negative (-1.972+0.041**loss*) for small values of *loss* (consistent with H2a.2) and positive for *loss* ≥ 1.972/0.041 = 48 (consistent with H2b.2). As we hypothesized, the negative effect of feedback with small losses can

be attributed to the *unlucky feeling* or attribution error by drivers (Billett and Qian 2008, Grossman and Owens 2012), which are known to lead to higher overconfidence and paying less attention to negative feedback. However, when the losses are significantly high, such losses are less likely to be due to external reasons or by chance. Therefore, drivers pay more attention to the feedback and as a result improve performance. Also note that the main effect of *loss* is negative, indicating that the deteriorating elements in performance are sticky, so they can be attributed to the demoralizing effect of negative feedback and carry into the following trips. Thus, in cases with large losses, it is crucial to provide feedback to compensate for the negative effect of loss itself. We note, however, that the high magnitude of losses (≥48 points) that is needed for feedback to become useful is quite rare, we find that it happens in about 3% of the cases.

Table 6: Effect of Gain and Loss on Performance

VARIABLES	(1) *oscore*
feedback	-1.972***
	(0.361)
gain×feedback	-0.059***
	(0.018)
gain	0.000
	(0.009)
loss×feedback	0.041**
	(0.016)
loss	-0.023**
	(0.009)
constant	77.876***
	(10.372)
R-squared	0.392

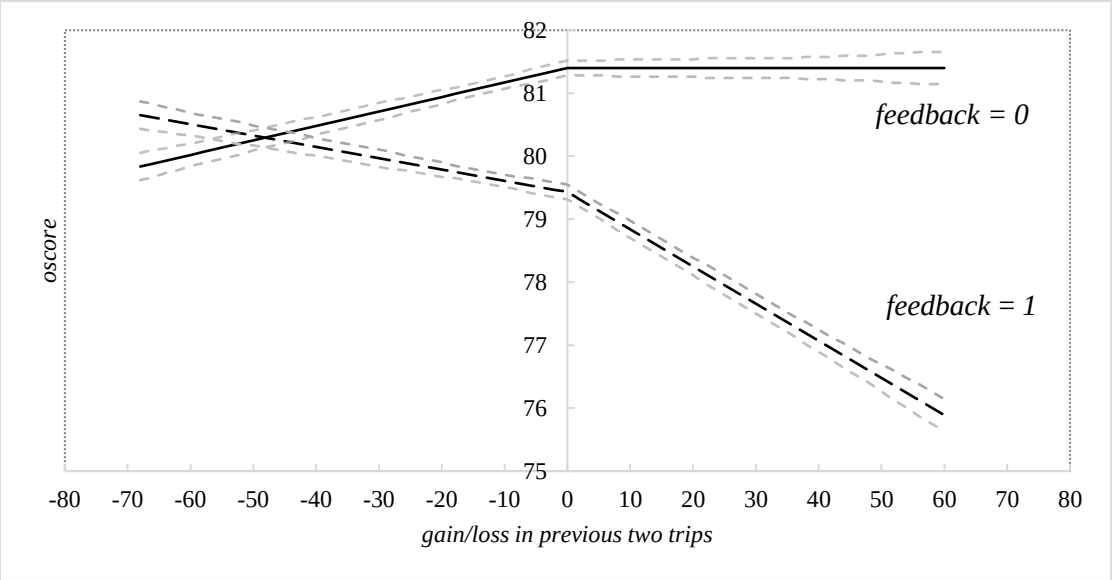
*Robust SEs in parentheses, *** p<0.01, ** p<0.05, * p<0.1, with individual, date, origin, and destination fixed effects and all controls (perrushhr, perdailyhr, pernighthr, freq, fbfreq, perphoneuse, ltodatescore, and loscore). SEs clustered for individuals, N = 80,503, estimation with the instrument, the first stage regression includes gain and loss variables as predictors.*

In summary, our results indicate that when the users observe an increase in score, their confidence might be increasing, which could lead to an increase in optimism bias, which is known to lead to risky behavior, resulting in a lower trip score in our case (H2b.1). When the users observe a decrease in score, their reaction depends on whether the decrease is small or large: A small decrease gets attributed to a potential error or bad luck, which leads users not to worry about their low scoring and even decrease their efforts (H2a.2). However, a large decrease in score shatters users' confidence and decreases their optimism bias, which leads to users working harder and driving more cautiously (H2b.2).

We illustrate these effects graphically in Figure 5. The solid line shows that, without feedback, gains do not help while losses hurt: namely, higher losses without feedback lead to lower performance. However, in the presence of feedback, higher gains lead to lower performance as illustrated by the black

dashed line. It further suggests that, in the presence of feedback, users incurring truly large losses tend to improve their performance, but this happens in a very small number of instances.

Figure 5: Gain vs. loss for feedback and no feedback condition *(gray color lines denote 95% CI)*



3.2.3. Feedback and Incentive Thresholds

To test Hypothesis 3, we construct a variable *nearThres* = 1 if a trip is taken when the user is "close" from below to the thresholds (70, 80, 90, and 95) and zero otherwise. We define "close" by including all the trips that are within *w* points of the thresholds. For instance, for threshold 70 and *w*=4, trips with *ltodatescore* within [66, 70] will have *nearThres* = 1. We present our results in Table 7, in which we interact *nearThres* with *feedback*. We test our results with different values of *w* = {3, 4, and 5}.

Table 7: Threshold Effect on Performance

	(1)	(2)	(3)
VARIABLES		*oscore*	
feedback	-3.085***	-3.488***	-3.529***
	(0.468)	(0.515)	(0.546)
nearThres× *feedback*	1.576***	2.185***	2.047***
	(0.507)	(0.555)	(0.580)
nearThres	-0.637***	-0.916***	-1.055***
	(0.242)	(0.306)	(0.322)
Constant	78.357***	78.413***	78.302***
	(10.705)	(10.684)	(10.579)
Controls	yes	yes	yes
R-squared	0.393	0.393	0.393
w	3	4	5

*Robust SEs in parentheses, *** p<0.01, ** p<0.05, * p<0.1, with individual, date, origin, and destination fixed effects and all controls (perrushhr, perdailyhr, pernighthr, freq, fbfreq, perphoneuse, ltodatescore, and loscore). SEs clustered for individuals. N = 80,503, estimation with the instrument.*

The three columns in Table 7 denote the estimation results when we change *w* from 3 to 5. All coefficients for the interaction term *nearThres*×*feedback* are positive and significant. Taking Column 2 as an example, if two users drive after receiving detailed feedback (i.e., *feedback=1*), one who is close to the threshold by four points from below will perform better by 1.269 (=2.185-0.916) points than the one who is not close to any threshold, everything else being equal. In terms of the standard deviation of *oscore* (i.e., 16.81), after feedback, users near the threshold will perform 7.54% (=1.269/16.81) in standard-deviation points better than the users whose *ltodatescore* are away from the threshold. Therefore, H3 is supported. We further test whether all thresholds are equally important and find that reviewing feedback close to the first discount threshold (i.e., 70) exerts a stronger effect compared with the other thresholds (we omit results for brevity).

3.3. Robustness Checks

a) Definition of feedback: As we argued earlier, we focus our analysis on the instances of detailed feedback (viewing multiple screens of the app to obtain feedback on driving behavior), rather than on the feedback obtained through simple push notifications. To further validate whether receiving a push notification is enough to affect users, we consider which operating system drivers use: according to Statista.com in Singapore about 60% of drivers use Android-based phones and about 40% use iOS-based phone, in our data we observe nearly 30% of the users with Android and rest with iOS. It is known (Accengage 2018) that, due to the different design of push notifications, Android users have about twice the percentage of opt-ins (91%) as iOS users (44%) and four times the interaction rate with notifications as Apple (iOS) users. The reason is that Android users automatically are opted in to receive notifications, and the notifications remain on the locked screen until dismissed, while with Apple devices, the users explicitly must opt in, and the notifications disappear from the locked screen after being unlocked once. We test whether a difference exists in performance between Android and Apple users, first using operating system, iOS and Android, dummies, then using an interaction model (please see Table 8 for detailed output with *OS=1* if the operating system of the user is iOS and zero otherwise). If feedback included in push notifications alone was impacting performance, we would see a difference between the Android and Apple users picked up by the OS coefficient. We conclude that push notifications alone do not exert a significant impact on performance, nor moderate the effect of feedback. This analysis supports our focus on detailed feedback. Admittedly, iOS and Android users can be very different along many dimensions, so our results are only suggestive.

b) Novelty effect: One could argue that the behavior that we observe in our data is partially due to learning. While becoming familiar with the app, the users may review feedback more frequently (novelty effect), which might vanish over time. Thus, the negative effect is observed only as users get used to the first few trips in the program. To disentangle the effect of learning, we code each of the first

10, 20, 30, and 40 trips with dummy variables, then we interact these variables with *feedback* and estimate our model. We summarize the results in Table E.4. We do not find any differences in the first 10 to 40 trips. We conclude that the negative effect of feedback persists beyond the first few trips; thus, our results are not subject to the novelty effect.

Table 8: Effect of Push Notifications on Trip Performance

	(1)	(2)
VARIABLES	*oscore*	
feedback	-2.429***	-2.273***
	(0.425)	(0.513)
feedback×OS		-0.548
		(0.994)
OS	-0.648	-0.467
	(0.591)	(0.696)
Constant	80.888***	79.257***
	(7.986)	(7.823)
Controls	yes	yes
Observations	79,639	79,639
R-squared	0.327	0.327

*Robust SEs in parentheses, *** p<0.01, ** p<0.05, * p<0.1, with date, origin, and destination fixed effects and all controls (perrushhr, perdailyhr, pernighthr, freq, fbfreq, perphoneuse, ltodatescore, and loscore). SEs clustered for individuals, estimation with instrument.*

c) High scores drive results: We noted that the trip scores are right-skewed, with a mean of 81.4 (refer to Table 1). An alternative explanation of the results could be that drivers are seeking feedback after receiving very high trip scores. Thereafter, it is likely that the drivers' score will go down, which can be attributed mistakenly to the impact of feedback. To address this possibility, we perform subsample analysis while excluding the high scores in addition to controlling for the *loscore*. We run our estimation only for the observations for which the *oscore* is below 95 (Models 1 and 2 in Table E.5) and below 90 (Models 3 and 4 in Table E.5). We observe that our results remain consistent even in the sub-sample, supporting the notion that our results are not driven by high scores. Notice that our estimates remain consistent, even when the sample size shrinks to fewer than 55,000 trips.

d) Effect of duration between feedback and the trip: One can argue that the time interval between a user reviewing detailed feedback and the next trip is an important variable: A trip taken long after a feedback review should exert less impact on performance. We introduce variable *fb2triptime* to code the time duration between feedback and trip starting time (in hours). We also include an interaction effect of *fb2triptime* and *feedback*. We report the results for this analysis in Table E.6. We do not find any evidence that the feedback effect is affected by the time interval between trip and feedback reviews. Our results for the effect of feedback on performance remain consistent. We observe an insignificant coefficient of the interaction term (Models 1 and 2 in Table E.6), i.e., the duration between feedback

and a trip exerts no effect on the performance regardless of whether or not the user reviews feedback.

4. Discussion and Implications

IoT devices and mobile connectivity in general increasingly enable user behavior to be tracked. Although extant conventional feedback literature provides ample support for the positive effect of conventional feedback on user behavior, little is known about the effect of immediate (close to real-time) feedback. UBI's main offering is to provide real-time or immediate evaluation of driving behavior to users and motivate them to improve it. We study the effect of such feedback in the case of automotive telematics, in which drivers may decide to study detailed driving-quality feedback to enable them to improve their driving behavior and earn insurance discounts.

Our results reveal several interesting findings. First, decision by drivers to review immediate feedback may not lead to improved behavior in their driving. On average, we find that reviewing feedback has a negative impact on the next trip's performance, namely, on the amount of time the subject spends driving over the speed limit. This negative effect is significant and can lead to a nearly one-year reduction in inter-accident time. To assess the negative impact that feedback could exert, consider the following back-of-the-envelope calculation. Let us assume that all drivers in Singapore adopt telematics devices, and the average score in our sample is representative of the driver population of Singapore. Our results indicate that reviewing *feedback* reduces the inter-accident time by 5.6%, i.e., the arrival rate of accidents will increase by 1.87% (=5.6/3%), as only on 1/3 of the trips the drivers look at their feedback on average. There were 7,690 accidents in Singapore in 2018 (MHA 2018); therefore, reductions in such scores due to feedback may result in 143 (=7,690*1.87%) additional accidents in Singapore. These numbers are not exaggerations of the potential impact of immediate feedback, as it has been estimated that 142 million drivers in the world would use such devices by 2023 (IHS Markit 2016).

Second, the type of feedback (e.g., positive or negative) matters because of drivers' overconfidence in their driving abilities. Namely, we find that viewing strong negative feedback (i.e., a significant deterioration in the driving score, happening in 3% of the cases) leads to improvement in short-term driving performance. On the other hand, viewing positive feedback (i.e., improved driving score) leads to further overconfidence and subsequent deterioration in the driving score. Therefore it might be prudent to provide only strong negative feedback to drivers. Third, drivers increase variation in their speed within a trip after reviewing detailed feedback, which results in a 2% increased probability of an accident. Fourth, the effect of the incentive threshold is very important, i.e., users who are close to the threshold from below react to feedback less negatively.

Our study relates to the studies exemplifying that sometimes, the information should not be shared. For instance, Lurie and Swaminathan (2009) find that providing feedback with high frequency can lead

to poorer performance. Therefore, it is important to withhold information for performance improvement. In our case, telematics firms are better off withholding feedback information when drivers do well or do a bit poorly.

Our study has limitations that are similar to the limitations of any research conducted with archival data. The first, and perhaps most important, limitation is generalizability of our results. We cannot observe drivers who are not participating in a telematics insurance program; however, we believe that these drivers would not be very different from the ones already in the program. In our context, users do not pay any setup costs. They can leave the app voluntarily if they do not find it useful. The discount obtained is on top of the no-claim discounts, and no penalties are assessed for poor scores. Maybe most importantly, there is no way to obtain data on users who do not download the app: Such data simply do not exist. Second, we have only one type of feedback, and it would have been interesting to study the effect of different types of feedback (e.g., social comparison). Third, our data prevent us from studying the role of demographics. It would be interesting to study the effects of age, education, and vehicle type in this context because conventionally, insurance companies consider these factors when calculating premiums. We use fixed effects for users to get around this issue. Finally, as our study duration is less than six months, we can analyze only short-term time-horizon driving. It is, indeed, important to study short-term performance in our application, as poor performance on any trip may lead to hazardous and potentially fatal consequences on the roads. However, in an ideal world, it also would be interesting to observe drivers' long-term performance. Methodologically, we are limited in the extent to which we can claim causal interpretation of our results since we do not assign treatment randomly to drivers. Our inference relies on an instrumental variable which, as always, can questioned. We also rely on a proprietary algorithm that relates feedback to driver over an aggregate score. There are always questions regarding interpretation of this total score by drivers who might not be unable to understand the composite scoring system.

Overall, our key finding calls for development of practical approaches to improve driving behavior, e.g., through changing or enhancing the way feedback is provided to drivers. One possible solution is nudging drivers to improve behavior through simple messages. For instance, in a follow up work, in a field experiment Choudhary et al. (2019) tested the efficacy of nudges that complement feedback by providing past-performance information to drivers while asking them to improve, and they found that effective forms of nudges are those that reference drivers' personal best and average performance levels. Choudhary et al. (2019) also conduct a lab experiment in which they demonstrate that nudges they propose work in a controlled laboratory setting and they stimulate increase in effort for subjects. Future work might include testing other types of nudges that may involve comparison of performance among drivers (social nudges) or even explicit recommendations to avoid certain routes/areas which are known for poor driving conditions. Further, Delgado et al. (2016) discuss a variety of approaches from legal

bans to technological solutions to nudges to reduce distracted driving. Yet, reducing traffic accidents remains remarkable difficult.

Acknowledgements

We thank Raxel Telematics for their collaboration, which included helping with data analysis and interpreting the results. We also would like to thank the Special Issue Editors, Associate Editor and two anonymous reviewers for providing us with constructive feedback on the previous versions of the manuscript.

References

Abadie A, Imbens GW (2011) Bias-Corrected Matching Estimators for Average Treatment Effects. *J. Bus. Econ. Stat.* 29(1):1–11.

Accengage (2018) *The 2018 Push Notification & in-App Message Benchmark*

Accenture (2018) UKI: Telematics—Passing Fad or Game-Changer? Retrieved (August 20, 2001), https://accntu.re/2zKQz0H.

Adams R, Almeida H, Ferreira D (2009) Understanding the relationship between founder-CEOs and firm performance. *J. Empir. Financ.* 16(1):136–150.

Anseel F, Beatty AS, Shen W, Lievens F, Sackett PR (2015) How Are We Doing After 30 Years? A Meta-Analytic Review of the Antecedents and Outcomes of Feedback-Seeking Behavior. *J. Manage.* 41(1):318–348.

Ashford SJ, Blatt R, VandeWalle D (2003) Reflections on the looking glass: A review of research on feedback-seeking behavior organizations. *J. Manage.* 29(6):773–799.

Ashford SJ, Tsui A (1991) Self-Regulation for Managerial Effectiveness : The Role of Active Feedback Seeking. *Acad. Manag. J.* 34(2):251–280.

Ashton RH (1990) Pressure and Performance in Accounting Decision Settings: Paradoxical Effects of Incentives, Feedback, and Justification. *J. Account. Res.* 28:148–180.

Automotive-Fleet (2016) Shaping Driver Behavior with Telematics. Retrieved (September 5, 2019), http://bit.ly/2VWmaZ1.

Azmat G, Bagues M, Cabrales A, Iriberri N (2018) What You Don't Know Can't Hurt You? A Field Experiment on Relative Performance Feedback in Higher Education. *Manage. Sci.* (9853):1–43.

Bandura A (1991) Social cognitive theory of self-regulation. *Organ. Behav. Hum. Decis. Process.* 50(2):248–287.

Barankay I (2012) Rank incentives: Evidence from a Randomized Workplace Experiment. *Bus. Econ. Public Policy Work. Pap.*

Berger J, Pope D (2011) Can Losing Lead to Winning? *Manage. Sci.* 57(5):817–827.

Billett MT, Qian Y (2008) Are overconfident CEOs born or made? Evidence of self-attribution bias

from frequent acquirers. *Manage. Sci.* 54(6):1037–1051.

Blader S, Gartendberg C, Prat A (2015) The Contingent Effect of Management Practices. *Columbia Bus. Sch. Res. Pap. No. 15-48*.

Buell RW, Norton MI (2014) Last-Place Aversion. *Q. J. Econ.*:105–149.

Burgers C, Eden A, van Engelenburg MD, Buningh S (2015) How feedback boosts motivation and play in a brain-training game. *Comput. Human Behav.* 48:94–103.

Cheema A, Bagchi R (2011) The Effect of Goal Visualization on Goal Pursuit: Implications for Consumers and Managers. *J. Mark.* 75(2):109–123.

Choudhary V, Shunko M, Netessine S, Koo S (2019) Nudging Drivers to Safety: Evidence from a Field Experiment. *SSRN Electron. J.*

Cramer J, Krueger AB (2016) Disruptive Change in the Taxi Business: The Case of Uber. *Am. Econ. Rev. Pap. Proc.* 106(5):177–182.

Croson R, Donohue K (2006) Behavioral Causes of the Bullwhip Effect and the Observed Value of Inventory Information. *Manage. Sci.* 52(3):323–336.

Dahlinger A, Ryder B (2018) The Impact of Abstract vs . Concrete Feedback Design on Behavior – Insights from a Large Eco-Driving Field Experiment. *Proc. CHI*:1–11.

Dehejia RH, Wahba S (2002) Propensity Score-Matching Methods for Nonexperimental Causal Studies. *Rev. Econ. Stat.* 84(1):151–161.

Delgado MK, Wanner KJ, McDonald C (2016) Adolescent Cellphone Use While Driving: An Overview of the Literature and Promising Future Directions for Prevention. *Media Commun.* 4(3):79.

DriverMetrics (2018) Maximising the Safety Benefits of Telematics. *http://bit.ly/2VswIQj*

Fischer M, Wagner V (2019) Effects of Timing and Reference Frame of Feedback: Evidence from a Field Experiment. *Ration. Compet.*:1–59.

Förster J, Higgins ET, Idson LC (1998) Approach and Avoidance Strength During Goal Attainment: Regulatory Focus and the "Goal Looms Larger" Effect. *J. Pers. Soc. Psychol.* 75(5):1115–1131.

Gjedrem WG (2018) Relative Performance Feedback: Effective or Dismaying? *J. Behav. Exp. Econ.* 74(July 2016):1–16.

Gnewuch U, Morana S, Heckmann C, Maedche A (2018) Designing Conversational Agents for Energy Feedback. *Proc. 13th Int. Conf. Des. Sci. Res. Inf. Syst. Technol. (DESRIST 2018)* (June).

Green PJ, Gino F, Staats B (2017) Shopping for Confirmation: How Disconfirming Feedback Shapes Social Networks. *Harvard Bus. Sch. Work. Pap.* (18–028).

GreenRoad (2019) Objective In-Cab Feedback. *https://greenroad.com/in-vehicle-feedback-2/*.

Grossman Z, Owens D (2012) An unlucky feeling: Overconfidence and noisy feedback. *J. Econ. Behav. Organ.* 84(2):510–524.

Hannan RL, Krishnan R, Newman AH (2008) The Effects of Disseminating Relative Performance

Feedback in Tournament and Individual Performance Compensation Plans. *Account. Rev.* 83(4):893–913.

Hasija S, Shen ZJM, Teo CP (2020) Smart City Operations: Modeling Challenges and Opportunities. *Manuf. Serv. Oper. Manag.* (Forthcoming):1–11.

IHS Markit (2016) Usage- Based Insurance Expected to Grow to Subscribers Globally by 2023 , IHS. Retrieved (September 14, 2018), https://bit.ly/2OvhLZa.

Ingenie (2019) Ingenie. Retrieved (May 10, 2019), https://www.ingenie.com/faqs.

Jenkins GD, Atul M, Gupta N, Shaw JD (1998) Are Financial Incentives Related to Performance? A Meta-Analytic Review of Em. *J. Appl. Psychol.* 83(5):777.

Kahneman D, Tversky A (1973) On the Psycology of Prediction. *Psychol. Rev.* 80(4):237–251.

Kivetz R, Urminsky O, Zheng Y (2006) The Goal-Gradient Hypothesis Resurrected: Purchase Acceleration, Illusionary Goal Progress, and Customer Retention. *J. Mark. Res.* 43(1):39–58.

Kluger AN, DeNisi A (1996) The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory. *Psychol. Bull.* 119(2):254–284.

Krueger N, Dickson PR (1994) How Believing in Ourselves Increases Risk Taking: Perceived Self-Efficacy and Opportunity Recognition. *Decis. Sci.* 25(3):385–400.

Lee TW, Locke EA, Phan SH (1997) Explaining the Assigned Goal-Incentive Interaction: The Role of Self-Efficacy and Personal Goals. *J. Manage.* 23(4):541–559.

Leverson B, Chiang KH (2014) Study of the Impact of a Telematics System on Safe and Fuel-efficient Driving in Trucks. *U.S. Dep. Transp.*

Luellen JK, Shadish WR, Clark MH (2005) Propensity Scores. *Eval. Rev.* 29(6):530–558.

Lurie NH, Swaminathan JM (2009a) Is timely information always better? The effect of feedback frequency on decision making. *Organ. Behav. Hum. Decis. Process.* 108(2):315–329.

Lurie NH, Swaminathan JM (2009b) Is Timely Information Always Better? The Effect of Feedback Frequency on Decision Making. *Organ. Behav. Hum. Decis. Process.* 108(2):315–329.

MHA (2018) On the Road with the Traffic Police. Retrieved (December 1, 2019), https://www.mha.gov.sg/hometeamnews/on-assignment/ViewArticle/on-the-road-with-the-traffic-police.

Mihm J, Schlapp J (2019) Sourcing Innovation: On Feedback in Contests. *Manage. Sci.* 65(2):559–576.

Moores TT, Chang JCJ (2009) Self-efficacy, overconfidence, and the negative effect on subsequent performance: A field study. *Inf. Manag.* 46(2):69–76.

MSIG (2018) Usage Based Private Motor. Retrieved (June 5, 2018), https://bit.ly/39l1fTv.

Nicholas JG, Gadirau R (1988) Speed Variance and its Influence on Accidents. *AAA Found. Traffic Saf.*

Nickell S (1981) Biases in Dynamic Models with Fixed Effects. *Econometrica* 49(6):1417.

NPR (2014) To Increase Productivity, UPS Monitors Drivers' Every Move. *https://n.pr/2VatawX*.

Peltzman S (1975) The Effects of Automobile Safety Regulation. 83(4):677–726.

PennMedNews (2018) Penn and CHOP Receive $1.84 Million to Study Ways to Curb Cell Phone Use while Driving. Retrieved (October 2, 2019), https://bit.ly/2Ztn1Qm.

Quddus M (2013) Exploring the Relationship Between Average Speed, Speed Variation, and Accident Rates Using Spatial Statistical Models and GIS. *J. Transp. Saf. Secur.* 5(1):27–45.

Renn R (2001) Development and Field Test of a Feedback Seeking, Self-Efficacy, and Goal Setting Model of Work Performance. *J. Manage.* 27(5):563–583.

Rivigo (2019) Safety is Central to Our Vision of Making Logistics Human. *https://www.rivigo.com/safety*. Retrieved (May 1, 2019), https://www.rivigo.com/safety.

Roels G, Su X (2014) Optimal Design of Social Comparison Effects: Setting Reference Groups and Reference Points. *Manage. Sci.* 60(3):606–627.

Rolim C, Baptista P, Duarte G, Farias T, Pereira J (2017) Impacts of Real-Time Feedback on Driving Behaviour: A Case-Study of Bus Passenger Drivers. *Eur. J. Transp. Infrastruct. Res.* 17(3):346–359.

Rosenbaum PR, Rubin DB (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70(1):41–55.

Roy MM, Liersch MJ (2013) I Am a Better Driver Than You Think: Examining Self- Enhancement for Driving Ability. *J. Appl. Soc. Psychol.* 43(8):1648–1659.

Schultz KL, Juran DC, Boudreau JW, McClain JO, Thomas LJ (1998) Modeling and Worker Motivation in JIT Production Systems. *Manage. Sci.* 44(12-part-1):1595–1607.

Schweitzer ME, Cachon GP (2000) Decision Bias in the Newsvendor Problem with a Known Demand Distribution: Experimental Evidence. *Manage. Sci.* 46(3):404–420.

Sharot T (2011) The optimism bias. *Curr. Biol.* 21(23):R941–R945.

Soleymanian M, Weinberg CB, Zhu T (2019) Sensor Data and Behavioral Tracking: Does Usage-Based Auto Insurance Benefit Drivers? *Mark. Sci.* 38(1):21–43.

Song H, Tucker AL, Murrell KL, Vinson DR (2018) Closing the Productivity Gap: Improving Worker Productivity Through Public Relative Performance Feedback and Validation of Best Practices. *Manage. Sci.* 64(6):2628–2649.

Staiger D, Stock J (1997) Instrumental Variables Regression with Weak Instruments. *Econometrica* 65(3):557–586.

Stern PC (1999) Information, Incentives, and Proenvironmental Consumer Behavior. *J. Consum. Policy* 22(4):461–478.

Tanishita M, van Wee B (2017) Impact of Vehicle Speeds and Changes in Mean Speeds on Per Vehicle-Kilometer Traffic Accident Rates in Japan. *IATSS Res.* 41(3):107–112.

Tiefenbeck V, Goette L, Degen K, Tasic V, Fleisch E, Lalive R, Staake T (2016) Overcoming Salience Bias: How Real-Time Feedback Fosters Resource Conservation. *Manage. Sci.* 64(3):1458–1476.

Toledo T, Lotan T (2006) In-Vehicle Data Recorder for Evaluation of Driving Behavior and Safety. *Transp. Res. Rec. J. Transp. Res. Board* 1953:112–119.

Toledo T, Musicant O, Lotan T (2008) In-Vehicle Data Recorders for Monitoring and Feedback on Drivers' Behavior. *Transp. Res.* Part C 16(3):320–331.

UPS (2016) *ORION : The algorithm proving that left isn't right*

Volpp KG, Troxel AB, Mehta SJ, Norton L, Zhu J, Lim R, Wang W, et al. (2017) Effect of Electronic Reminders, Financial Incentives, and Social Support on Outcomes After Myocardial Infarction: The HeartStrong Randomized Clinical Trial. *JAMA Intern. Med.* 19104:1–9.

E-companion

Section E.1: Linear (OLS) Regression

We report results for the estimation of Equation (1) in Table E.1. In Model 1, we estimate the coefficients using only the control variables, and we find expected directions for their effects. We observe that performance is associated negatively with night hours (-0.145, p<0.01), while it is associated positively with daily hours driven (0.148, p<0.01). Furthermore, we find a positive association of $oscore_{ijt}$ with the past trip score, as well as with to-date score. We find a positive association of phone usage with the dependent variable (0.034, p<0.01). This may be due to drivers primarily using map applications (such as Google Maps and Waze) to identify optimal routes. Next, Model 2 includes the first treatment variable (*feedback*), along with the date, individual, origin, and destination fixed effects. We observe that feedback is associated negatively with users' performance during the next trip. The coefficient of $feedback_{ijt}$ is significant and negative (in Model 2, the coefficient is -0.837, p<0.01), indicating that once users review their feedback, they perform worse in the next trip by 0.837 points, contrary to what we predicted in H1a, but consistent with H1b.

Table E.1: Effect of Feedback on Next Trip Score

DV=*oscore*	(1)	(2)
feedback		-0.837***
		(0.160)
perdailyhr	0.148***	0.147***
	(0.003)	(0.003)
pernighthr	-0.145***	-0.147***
	(0.005)	(0.005)
freq	0.206	0.195
	(0.129)	(0.129)
fbfreq	0.006	0.017
	(0.010)	(0.010)
perphoneuse	0.034**	0.034**
	(0.013)	(0.013)
ltodatescore	0.003	0.004
	(0.033)	(0.033)
loscore	0.074***	0.073***
	(0.005)	(0.005)
Constant	79.073***	78.951***
	(11.113)	(10.947)
R-squared	0.394	0.395

*Robust SEs in parentheses, *** p<0.01, ** p<0.05, * p<0.1, with individual, date, origin, and destination fixed effects and all controls (perrushhr, perdailyhr, pernighthr, freq, fbfreq, perphoneuse, ltodatescore, and loscore). SEs clustered for individuals, N = 80,503.*

Section E.2: Assessing the Impact From Feedback on Driving Performance Using Matching

Propensity scoring (for details, see Rosenbaum and Rubin 1983) is the standard way to match a variable in a control group to statistically equivalent subjects within a treatment group in the absence

of a random assignment to groups. This method has been employed successfully by many researchers as a way to select an unbiased control group when studying a treatment's effect. The key idea of matching, in our case, is that trips that belong to two different groups—such as treatment and control—are, nevertheless, comparable if they have similar propensity scores (Luellen et al. 2005).

We compare treated trips to control trips that are best matched based on observed parameters to achieve the best possible balance on covariates across the feedback trips and the matched non-feedback trips, such that we have the closest estimate of the marginal effect of *feedback* on *oscore*. We consider two approaches to specify how we create an appropriate control set for the treatment trips: (a) propensity score matching and (b) nearest-neighbor matching. Once we have the control group of non-feedback trips for every treatment trip, we estimate the average treatment effect (ATE) of feedback based on the mean score of the treatment trips vs. the mean score of the control trips. First, we utilize matching based on propensity scores, which typically are calculated as the predicted probabilities of group membership, such as the probability of seeking feedback, based on an appropriate set of observable factors. Thus, propensity scores can be calculated based on an estimation of a binary response model, such as *probit*. We estimate the probability of feedback seeking using control variables from the previous trip (lagged control variables) and time-invariant factors, i.e., origin and destination. We include individual and date-fixed effects as follows:

$$Pr\left(feedback_{ijt} = 1\right) = \Phi\left(\alpha_0 + \beta\ Controls_{ij-1t} + date_t + individual_i + orig_{j-1} + des_{j-1}\right). \quad (3)$$

Table E.2: Probit Model With Feedback as a Binary Outcome Variable

VARIABLES	*feedback*
lperdailyhr	-0.000
	(0.000)
lpernighthr	0.005***
	(0.000)
lfreq	0.057***
	(0.017)
lfbfreq	0.006***
	(0.001)
lperphoneuse	-0.000
	(0.001)
ltodatescore	0.002
	(0.003)
loscore	-0.002***
	(0.000)
Constant	-0.203
	(1.086)

*Robust SEs in parentheses, *** p<0.01, ** p<0.05, * p<0.1, with individual, date, origin, and destination fixed effects and all controls (perrushhr, perdailyhr, pernighthr, freq, fbfreq, perphoneuse, ltodatescore, and loscore). SEs clustered for individuals, N =79,648*

Prior research suggests that a small set of matching variables can provide reasonably good predictive power (Dehejia and Wahba 2002). We present the results from the probit model (Equation 3) in Table E.2. Note that we use prefix "*l"* to denote lagged variables. Of the seven variables, four are significant at a $p<0.01$ level. A higher percentage of night hours, trip frequency, and feedback frequency are associated with higher feedback seeking before the next trip. However, a higher previous-trip score has a negative association with feedback viewing (coefficient -0.002, $p<0.01$). We note a smaller number of observations than the original dataset, considering that 16 users either did not review any feedback or reviewed detailed feedback nearly after each trip.

Given a trip with detailed feedback and its predicted probability of being a treatment trip as determined by the *probit* estimation, we take the most similar non-feedback trip to be the one that shows the numerically closest *ex ante* predicted probability of feedback, with a replacement. Once we determine the best match for each treatment trip, we compute the mean score for each group and perform a difference of means t-test, just as before. We show these results in the first column of Table E.3. The first column shows the size of the ATE, which is the difference of the mean of the scores between the feedback and non-feedback trips. A matched sample based on propensity scores suggests that feedback induces a negative impact on the score of the next trip (coefficient-1.475, $p<0.01$). The effect size is quite similar to the coefficient obtained without matching: The OLS slightly underestimated the coefficient when we compare it to the ATE provided by propensity score matching.

Recent seminal work (Abadie and Imbens 2011) has questioned the large-sample properties of propensity score matching estimators. They propose a set of nearest-neighbor (NN) matching estimators for the treated units that provide each unit with a control group of NN $\geq$ 1 untreated (control) observations whose covariates are most similar to those of the treated unit. We set NN to one in the results reported below, i.e., each trip with feedback is matched to at least one nearest neighbor with replacement.

We use the same covariates for matching that we employed for *probit* estimation in Table E.2. We report results of matching in Columns 2 to 6 in Table E.3, in which the first row shows the effect of feedback on *oscore.* We employ a bias-adjusted estimator using all continuous control variables as prescribed in extant literature to obtain a consistent estimator (Abadie and Imbens 2011). We calculate robust standard errors, and we observe that our results are consistent across models. We also try exact matching of variables (date, individual, and both) and obtain consistent results that further reinforce earlier analysis (refer to Models 4-6). Note that we cannot use robust standard error when we are conducting exact matching using both individual and date variables (Column 6), as the number of observations drops significantly (less than 8,000). Our estimated effect size using OLS is within the range of the results obtained using matching.

Table E.3: Results From Matching: Effect of Detailed Feedback on *oscore*

	Propensity score matching	Nearest-Neighbor Matching				
	(1)	(2)	(3)	(4)	(5)	(6)
feedback (1 vs. 0)	-1.820***	-1.888***	-0.981***	-0.989***	-0.594***	-0.816***
	(0.178)	(0.245)	(0.186)	(0.177)	(0.136)	(0.137)
Observations	79,648	79,648	80,503	77,881	80,503	55,710
Matched Treatments	26,042	26,042	26,042	26,012	26,042	21,236
Matching Variables	-	propensity score	all control variables			
Bias-adj.	-	-	yes	yes	yes	yes
Robust std. error	yes	yes	yes	yes	yes	no
Exact matching Variables	-	-	-	individual	date	date and individual

*Standard errors reported in parentheses, *** p<0.01, ** p<0.05, * p<0.1*

Section E.3: Additional Tables

Table E.4: Robustness Test for Novelty Effect

	(1)	(2)	(3)	(4)
VARIABLES		*oscore*		
feedback	-2.500***	-2.516***	-2.508***	-2.529***
	(0.383)	(0.386)	(0.390)	(0.394)
first10trip×feedback	-0.497			
	(1.184)			
first10trip	0.318			
	(0.642)			
first20trip×feedback		0.101		
		(0.748)		
first20trip		-0.396		
		(0.455)		
first30trip×feedback			0.015	
			(0.644)	
first30trip			0.037	
			(0.366)	
first40trip×feedback				0.214
				(0.603)
first40trip				-0.097
				(0.328)
Constant	78.756***	78.789***	78.704***	78.698***
	(10.448)	(10.570)	(10.510)	(10.559)
R-squared	0.393	0.393	0.393	0.393
Controls	yes	yes	yes	yes
for	first 10 trips	first 20 trips	first 30 trips	first 40 trips

*Robust SEs in parentheses, *** p<0.01, ** p<0.05, * p<0.1, with individual, date, origin, and destination fixed effects and all controls (perrushhr, perdailyhr, pernighthr, freq, fbfreq, perphoneuse, ltodatescore, and loscore). SEs clustered for individuals. N = 80,503, estimation with instrument.*

Table E.5: Regression Results for High Trip Scores Sub-Sample

	(1)	(2)
VARIABLES	*oscore*	
feedback	-2.386***	-2.250***
	(0.370)	(0.369)
Constant	71.751***	72.215***
	(3.111)	(3.143)
Observations	53,633	54,239
R-squared	0.264	0.265
Controls	yes	yes
oscore range	<90	<95

*Robust SEs in parentheses, *** p<0.01, ** p<0.05, * p<0.1, with individual, date, origin, and destination fixed effects and all controls (perrushhr, perdailyhr, pernighthr, freq, fbfreq, perphoneuse, ltodatescore, and loscore). SEs clustered for individuals, estimation with instrument.*

Table E.6: Effect of Time to Feedback Duration on Performance

	(1)
VARIABLES	*oscore*
feedback	-1.594***
	(0.540)
fb2triptime×feedback	-0.003
	(0.002)
fb2triptime	-0.000
	(0.000)
Constant	80.010***
	(10.525)
Controls	yes
R-squared	0.392

*Robust, standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1, with individual, date, origin, and destination fixed effects and all controls (perrushhr, perdailyhr, pernighthr, freq, fbfreq, perphoneuse, ltodatescore, and loscore). SEs clustered for individuals, estimation with instrument, N = 80,503.*

Table E.7: First-stage Instrumental Variable Regression

	(1)
VARIABLES	*feedback*
perdailyhr	-0.000**
	(0.000)
pernighthr	-0.001***
	(0.000)
freq	0.029***
	(0.004)
fbfreq	0.008***
	(0.000)
perphoneuse	0.000**
	(0.000)
ltodatescore	-0.000
	(0.001)
loscore	-0.000
	(0.000)
$\widehat{feedback}$	1.020***
	(0.010)
Constant	-0.225
	(0.212)
R-squared	0.482

*Robust, standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1, with individual, date, origin, and destination fixed effects. SEs clustered for individuals, N = 80,503.*